Divergence in Expression between Duplicated Genes in Arabidopsis

Eric W. Ganko, Blake C. Meyers, and Todd J. Vision

*Department of Biology, University of North Carolina at Chapel Hill and †Department of Plant and Soil Science, University of Delaware

New genes may arise through tandem duplication, dispersed small-scale duplication, and polyploidy, and patterns of divergence between duplicated genes may vary among these classes. We have examined patterns of gene expression and coding sequence divergence between duplicated genes in Arabidopsis thaliana. Due to the simultaneous origin of polyploidy-derived gene pairs, we can compare covariance in the rates of expression divergence and sequence divergence within this group. Among tandem and dispersed duplicates, much of the divergence in expression profile appears to occur at or shortly after duplication. Contrary to findings from other eukaryotic systems, there is little relationship between expression divergence and synonymous substitutions, whereas there is a strong positive relationship between expression divergence and nonsynonymous substitutions. Because this pattern is pronounced among the polyploidy-derived pairs, we infer that the strength of purifying selection acting on protein sequence and expression pattern is correlated. The polyploidy-derived pairs are somewhat atypical in that they have broader expression patterns and are expressed at higher levels, suggesting differences among polyploidy- and nonpolyploidy-derived duplicates in the types of genes that revert to single copy. Finally, within many of the duplicated pairs, 1 gene is expressed at a higher level across all assayed conditions, which suggests that the subfunctionalization model for duplicate gene preservation provides, at best, only a partial explanation for the patterns of expression divergence between duplicated genes.

Introduction

Functional divergence among duplicated genes is among the most important sources of evolutionary innovation in complex organisms. Population genetic theory suggests that most duplicate gene pairs revert to single copy on a short evolutionary timescale and that those pairs that are retained are likely to have at least partially diverged in function (Lynch and Conery 2000). Two indicators of functional divergence that are amenable to systematic study are changes in protein sequence and alterations in the timing, location, and relative number of gene transcripts.

A number of studies in recent years have measured the divergence between duplicate genes in their spatiotemporal profiles of gene expression and compared this with proxies of duplication age, such as the number of synonymous substitutions between the 2 aligned protein coding sequences. Using spotted cDNA data from yeast, Gu et al. (2002) found a significant positive relationship between the divergence in expression and the estimated number of synonymous (\(K_S\)) and nonsynonymous (\(K_A\)) substitutions per site, though the relationship held only for \(K_A \leq 0.3\). This contradicted an earlier, more limited, study where the weak relationship between correlation in expression and protein sequence divergence was not significant (Wagner 2000). In agreement with Gu et al. (2002), Makova and Li (2003) showed that \(K_S\) and \(K_A\) are positively related to the divergence in expression among different tissues in human duplicate gene pairs for \(K_A \leq 0.2\). Considerable divergence in expression had occurred in about 50% of the pairs even when there was very little sequence divergence (\(K_S < 0.1\)). This suggests that gene pairs diverge rapidly in their spatiotemporal pattern of expression at or near the time of inception and continue to diverge steadily thereafter.

Studies such as those cited above depend on the assumption that \(K_S\) and to a lesser degree \(K_A\), can be treated as proxies for time. Although it is well known that this is not strictly true (e.g., Zhang et al. 2002), it is generally assumed that variation in substitution rates will only obscure whatever trends may exist between expression divergence and time but would not be positively misleading. Furthermore, it is assumed that variation in the substitution rate among gene pairs is not related to variation in the rate of expression divergence. Contrary to this assumption, recent studies of multiple pairs of orthologous genes between closely related Drosophila species (Nuzhdin et al. 2004; Lemos et al. 2005) and between distantly related Caenorhabditis elegans and Caenorhabditis briggsae (Castillo-Davis et al. 2004) suggest that divergence in protein sequence and gene expression may be coupled. These findings raise the question of whether divergence in protein sequence and gene expression are also coupled in paralogous gene pairs.

Arabidopsis thaliana is an ancient polyploid and thus has large numbers of “segmental duplicates” of near-simultaneous age (Vision et al. 2000; Zhang et al. 2002; Blanc et al. 2003). For this reason, and because of the broad and deep expression data available (Meyers, Tej, et al. 2004; Schmid et al. 2005), Arabidopsis is an attractive model for studying the relationship between substitution rate and expression divergence. Previously, Haberer et al. (2004) estimated that two-thirds of duplicate gene pairs had divergent expression in Arabidopsis. Casneuf et al. (2006) concluded that nonsegmentally duplicated pairs diverge more rapidly than segmentally duplicated pairs and also that the rate of expression divergence is somewhat dependent upon inferred gene function.

In addition to segmental duplicates, we recognize 2 other categories of duplicated gene pairs in this study. “Tandem duplicates” are closely adjacent to one another in the genome and most likely arise through a mechanism of unequal crossing over. “Dispersed duplicates” are found neither adjacent to each other in the genome nor within homeologous chromosome segments. A number of different mechanisms may directly create dispersed pairs of duplicated genes in eukaryotes (Katju and Lynch 2003). Some dispersed pairs may also be due to relocation of one copy of a tandem or segmental duplication through a secondary
chromosomal rearrangement. Recent work suggests that the evolutionary dynamics of duplicate gene retention and functional divergence may differ substantially among these categories (reviewed in Vision 2005; Freeling and Thomas 2006).

Gene expression divergence may play an important role in the preservation of duplicated genes. The probability of preserving both duplicates is quite small if preservation depends solely on “neofunctionalization,” the appearance of a mutation that modifies one copy to perform a new, adaptively favored, function. More frequently, a deleterious mutation would be expected to appear first that would lead to nonfunctionalization or loss of one of the duplicate copies (Walsh 1995). An alternative mechanism by which both copies of a pair can be prevented from undergoing nonfunctionalization is via “subfunctionalization” or the acquisition of complementary degenerative mutations in the 2 copies that partition some of the functionality of the ancestral gene into its 2 descendents (Force et al. 1999). One way in which subfunctionalization can occur in organisms with complex gene regulation is through complementary changes in transcription factor–binding sites that result in differential expression in the 2 copies. In the simplest formulation of the model, there would be at least 1 condition or cell type in which each copy is expressed and the other is not.

Lynch et al. (2001) have shown that the probability that a duplicate pair is retained through subfunctionalization versus neofunctionalization depends on several factors, including population size and the frequency of recombination between the pair. Because subfunctionalization does not depend on the occurrence of a favorable mutation, it is more likely than neofunctionalization to explain the retention of duplicate genes in organisms with small population sizes, such as multicellular plants. Additionally, tight linkage facilitates subfunctionalization because of population genetic processes acting during fixation of the new duplication. Because tandem duplicates are in close linkage, this leads to the prediction that subfunctionalization should be a more frequent mechanism for the retention of tandem duplicates than for other kinds of duplicated pairs. We can test this prediction by comparing young tandem pairs with young dispersed pairs.

Two Arabidopsis gene expression data sets are used in this study. One derives from Massively Parallel Signature Sequencing (MPSS, Brenner et al. 2000) described more fully in Meyers, Tej, et al. (2004). MPSS provides a digital signature of mRNA abundance for each gene, similar to that obtained using Serial Analysis of Gene Expression, and has 2 attractive features for this study. 1) mRNA abundance can be reliably quantified even at low levels because there is essentially no noise. 2) The expression measurements for closely related duplicated genes are not subject to cross-hybridization artifacts, which is a major impediment to studying divergence among gene family members using hybridization-based gene expression assays (Xu et al. 2001).

The second data set is the Arabidopsis Development Atlas (ADA, Schmid et al. 2005), which contains triplicate expression estimates for ~80% of known Arabidopsis genes across 79 different tissues and developmental time points using the Affymetrix ATH1 microarray. In contrast to the MPSS data set, which consists of 5 coarsely sampled mRNA libraries, the ADA data set allows us to detect relatively fine anatomical and developmental differences between expression profiles. However, in order to minimize the effects of cross-hybridization in the ADA data set, it was necessary to exclude array probes that matched more than 1 gene; as a result, there is relatively poor sampling of duplicates with small amounts of sequence divergence.

Materials and Methods

Sequence Analysis of Duplicate Pairs

For this study, we focus on gene families containing only 2 members. Because divergence between such pairs is (relatively) free of influence by functional overlap with other gene family members, the divergence in expression pattern between them is relatively straightforward to interpret. However, we cannot exclude the possibility that extinct gene family members may have influenced the divergence of the pairs we study.

Pairs of duplicated genes were selected as follows. An all-by-all BlastP (Altschul et al. 1997) with default parameters was performed with all annotated Arabidopsis protein sequences from TIGR version 3.0. The Blast output was processed by TribeMCL (Enright et al. 2002), with an expansion parameter of 5, to cluster each gene uniquely into one gene family. Each pair was classified as either tandem (T), segmental (S), or dispersed (D). If the 2 members of a pair were within 20 annotated genes of each other on the same chromosome, they were assigned to class T. If they were identified as anchors of a segmental homology using fluorescent in situ hybridization (Calabrese et al. 2003), they were assigned to class S. The remaining pairs were assigned to class D.

Each pair of protein sequences was aligned using ClustalW (Thompson et al. 1994) using default parameters. The DNA sequences were aligned codon-by-codon, using the protein sequences as guides, and per-site synonymous (Ks) and nonsynonymous (Ka) substitution rates were calculated using PAML (Yang 1997). Pairs with Ks ≥ 1 were discarded. Codon bias, measured as the effective number of codons (ENC), was calculated for each gene pair using CodonW (Wright 1990) to test whether highly biased genes were present in the data set that would lead to anomalously low estimates of the number of synonymous substitutions between duplicate genes (Friedman and Hughes 2001). Because we wished to restrict our study to functional genes, annotated pseudogenes were ignored. In addition, we examined each gene pair for evidence of premature truncation. Less than 10% of gene pairs differ in size by more than 25% (relative to the smaller of the pair). These pairs are evenly distributed among the duplication categories.

Measuring Gene Expression Using MPSS Data

We analyzed transcript abundance in 5 CDNA libraries derived from callus, inflorescence, leaf, root, and silique tissues harvested under standard growth conditions from A. thaliana accession Col-0. Details of plant growth, library
construction, and the analysis of the MPSS sequence signatures obtained are reported in Meyers, Tej, et al. (2004) and Meyers, Vu, et al. (2004). Signatures were obtained using the “classic” MPSS method. Although the total number of signatures obtained from each library ranged from approximately 1.7 to 3.6 million, all signatures are normalized to transcripts per million (TPM) (Meyers, Tej, et al. 2004). These data are available from http://mpss.udel.edu/at/.

Measuring Gene Expression Using Affymetrix Data

We also analyzed expression data from the ADA (ExpressionSet_ME00319 at ftp://ftp.arabidopsis.org/), which contains triplicate expression estimates for each of 79 tissues and developmental time points using the Affymetrix ATH1 chip (Schmid et al. 2005). The ATH1 chip has 22,746 probes representing >80% of known Arabidopsis genes. We matched each Affymetrix probe to the TAIR Arabidopsis genome annotation (ftp://ftp.arabidopsis.org/home/tair/Microarray/old/affy8k_array_elements-2004-06-01.txt) and excluded from analysis any probes that matched multiple TAIR genes. The mean signal value of each triplicate was calculated for each probe, under each condition, only when all 3 replicates had a “present” call. All other instances were treated as absent with a signal value of 0.

Analysis of Expression Divergence between Duplicates

We employed a number of different measures of the divergence between 2 expression profiles, including $A_E$, average level of expression per library; $I_L$, proportion of shared libraries; and $N_L$, number of libraries where at least 1 copy of a pair is expressed (see Results). The normalized Manhattan distance $d_N$ calculates the distance between the expression vectors of genes $i$ and $j$:

$$d_N = \frac{1}{2} \sum_{k=1}^{5} \left| \frac{x_{i,k}}{\sum_{k=1}^{5} x_{i,k}} - \frac{x_{j,k}}{\sum_{k=1}^{5} x_{j,k}} \right|,$$

where $x_{i,k}$ is the expression of gene $i$ in library $k$ and each element of the vector is standardized by $\sum_{k=1}^{5} x_{i,k}$, the total expression of that gene over all libraries. It is more sensitive to the proportion of libraries that differ among the genes than an alternative measure based on Euclidean distance.

Another measure of the similarity in the expression profiles of 2 genes is the Pearson product-moment correlation, $r_P$; this measure is more useful for the ADA data set because there are an insufficient number of libraries in the MPSS data set to obtain a reliable correlation estimate. Note that both measures are only sensitive to differences in the relative abundance of transcripts among libraries and not to absolute differences in abundance between 2 pairs of genes. Following Gu et al. (2002) and Makova and Li (2003), we used a log($1 + r_P$)/($1 - r_P$) transformation for the ADA data set. However, we found an arcsin $\sqrt{r_P}$ transformation to yield much more normally distributed residuals for the MPSS data set, and so we report these results using this transformation. The qualitative conclusions are not dependent on the transformation chosen.

In order to compare the observed statistics with those expected for gene pairs that had diverged to the point of randomness, the same statistics were calculated for 100 MPSS and ADA data sets in which gene identities and expression profiles had been randomly paired.

Each measure of divergence was compared with $K_A$ and $K_S$ by linear regression for both observed and shuffled data sets. A Bonferroni-corrected $P = 5.2 \times 10^{-4}$ (corresponding to an alpha value of 0.05) was used as a threshold of significance to account for multiple testing.

Measuring Symmetry of Expression Divergence

We used different tests to determine whether duplicated genes differed in expression level within a library. For the MPSS data set, we compared the higher signal with a 95% binomial confidence interval around the lower signal value. For the ADA data set, the higher signal was compared with the 95% prediction limit (Sokal and Rohlf 1995) for a sample size of 3 fitted to the lower signal value.

Results

Divergence between gene pairs from 2-member gene families is relatively free of influence by functional overlap with other gene family members, simplifying the interpretation of expression pattern divergence between them. Clustering of the pairwise matches obtained from an all-by-all sequence similarity search of the predicted Arabidopsis proteome yielded 984 gene families of size 2. Of these, 503 pairs were estimated to have an average level of synonymous divergence among sites $K_S < 1$ and were retained for analysis (supplementary table 1, Supplementary Material online). All genes in this set had an ENC greater than 35, indicating that synonymous substitutions were not constrained by codon usage. A comparison of the chromosomal positions of the members of each pair relative to one another and to duplicated genomic segments within Arabidopsis allowed us to subdivide all the duplicated pairs into 65 tandem, 152 segmental, and 286 dispersed pairs (supplementary table 1, Supplementary Material online). Of the 152 segmental pairs, all but 5 were present in the data set of segmentally duplicated pairs of Blanc et al. (2003), and all those that were present were classified as having occurred during the most recent large-scale duplication event 40–80 MYA (Lynch and Conery 2000; Vision et al. 2000; Simillion et al. 2002; Blanc et al. 2003; Bowers et al. 2003). We treat these duplicates as having arisen simultaneously and interpret the variability in $K_S$ and $K_A$ among these pairs to be due to variability in the rate of substitution as in Zhang et al. (2002).

Transcript Abundance in the Data Set Relative to Other Genes

Levels of expression were measured by MPSS for each gene in 5 different libraries (callus, flower, leaf, root, and silique). Of the 503 gene pairs, 11 had a total expression level of less than 5 TPM summed across all 5 libraries in
both members of the pair and were discarded. In all, 492 gene pairs remained for study in the MPSS data set and were subdivided into 63 tandem, 149 segmental, and 280 dispersed pairs. The 984 genes represented by the 492 pairs had median expression levels ranging from 11.5 to 15 TPM and maximum expression levels ranging from 3,000 to 5,000 TPM among the libraries. Using a threshold of at least 5 TPM to declare that a gene is expressed in a library, there is an unusual bimodality in the pattern of gene pair expression observed. The majority of genes are expressed in either less than 2 libraries or constitutively in all 5. The distribution is similar among all Arabidopsis genes, though there are somewhat more constitutively expressed genes among the 984 in the duplication data set (table 1A). In over two-thirds of the pairs (345), each copy is expressed in at least 1 library. Interestingly, these jointly expressed pairs are more frequent in the segmental class (76.5%) than in the dispersed (67.5%) and tandem (66.6%) classes, respectively (Fisher’s exact test, two-tailed: \( P < 2 \times 10^{-16} \)).

We performed a similar analysis on the ADA data set. To account for potential cross-hybridization problems in these data, gene pairs that could not be unambiguously matched to probes on the ATH1 chip were discarded. The 325 remaining pairs there were subdivided into 185 dispersed, 110 segmental, and 30 tandem duplicate pairs. Using the present call as a threshold for minimal expression within the ADA data set, we observed a median signal of 231, with levels ranging from 5 to 37,000, across all 79 libraries. Unlike the MPSS data set, we did not observe a bimodal distribution of expression among duplicate pairs but rather an abundance of pairs with broad expression in many libraries (table 1B). Approximately one-third (32%) of duplicate genes were expressed in all 79 libraries and over two-thirds were expressed in >70 tissue and developmental libraries. Only 11 duplicate gene pairs (2%) lacked a significant signal value in all 79 libraries, much lower than the 12% of all genes lacking a single significant value (table 1B).

### Transcript Abundance and Sequence Conservation

We then compared the total expression of a pair, \( T_E \), with the rate of synonymous (\( K_S \)) and nonsynonymous substitution (\( K_A \)) between pairs. As with all the subsequent analyses relating sequence divergence to some measure of expression, the data are quite noisy and the linear trends we identify only explain a portion of the variance. Nonetheless, a number of striking patterns are evident.

Summing across all 5 libraries in the MPSS data, we found that \( T_E \) declined precipitously with \( K_A (\log(T_E), \ P < 6.72 \times 10^{-15}; \text{seen in fig. 1A}) \). The decline was approximately twice as steep among segmental pairs as in dispersed and tandem pairs (interaction between \( K_A \) and duplication type; \( P = 0.024 \), and the negative relationship was significantly different from 0 for the segmental and dispersed duplication types when considered separately. The median \( K_A \) for pairs with a total expression level greater than 1,000 TPM was 0.05 versus 0.13 for pairs with expression levels of 1,000 TPM or less. In contrast, the total expression level was weakly related to \( K_S \) only within segmental duplicates and not significant after correcting for multiple tests. A significant negative relationship was also observed between \( T_E \) and \( K_A (\log(T_E), \ P < 2.2 \times 10^{-16}) \) for all gene pairs in the ADA data (fig. 1B) and also within the segmental and dispersed duplication types separately. In agreement with the MPSS data, no significant relationship was observed between \( T_E \) and \( K_S \), \( T_E \) was, on average, higher than that found in the randomized data only for gene pairs with \( K_A \) less than \( \sim 0.1 \) (fig. 1).

\( T_E \) conflates the breadth of expression and the average level of expression within those libraries in which the genes are expressed, so we separately quantified the breadth of expression by counting the number of libraries (\( N_L \)) in which at least one of the copies is expressed. We found that \( N_L \) declined with increasing \( K_A \) independent of duplication class in both data sets (fig. 2) and showed no significant relationship to \( K_S \) (results not shown).

We examined abundance independently of breadth by measuring the average level of expression (\( A_E \)) among those libraries in which expression was substantially different from 0 (\( \geq 5 \) TPM in the MPSS data set and a present call for all 3 replicates in the ADA data set). With MPSS, average expression showed a significant interaction between \( K_A \) and duplication type even when \( N_L \) was included in the model (\( P = 2.3 \times 10^{-5} \)). The interaction was due to a significant negative relationship between \( A_E \) and \( K_A \) among segmental and dispersed, but not tandem, pairs. The slope was approximately 3 times greater for segmental than dispersed pairs (fig. 3A). We also observed a significant interaction between \( K_A \) and \( A_E \) in segmental and dispersed pairs in the ADA data, where the slope of segmental pairs was approximately 2 times greater than dispersed pairs (fig. 3B). In both data sets, no significant interaction was detected between \( K_S \) and \( A_E \). Thus, considering the combined results for \( T_E, N_L, \text{and } A_E \) in segmental pairs, it appears that both the breadth of expression and the absolute level of expression decline most strongly with the rate of nonsynonymous substitution.

### Table 1

<table>
<thead>
<tr>
<th>A. MPSS</th>
<th>Genes in pairs (%)</th>
<th>All genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>153 (16)</td>
<td>4160 (23)</td>
</tr>
<tr>
<td>1</td>
<td>124 (13)</td>
<td>2643 (15)</td>
</tr>
<tr>
<td>2</td>
<td>73 (7)</td>
<td>1727 (10)</td>
</tr>
<tr>
<td>3</td>
<td>93 (9)</td>
<td>1777 (10)</td>
</tr>
<tr>
<td>4</td>
<td>109 (11)</td>
<td>1930 (11)</td>
</tr>
<tr>
<td>5</td>
<td>432 (44)</td>
<td>5612 (31)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. ADA</th>
<th>Genes in pairs (%)</th>
<th>All genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11 (2)</td>
<td>2693 (12)</td>
</tr>
<tr>
<td>1–10</td>
<td>38 (6)</td>
<td>3063 (14)</td>
</tr>
<tr>
<td>11–20</td>
<td>26 (4)</td>
<td>1403 (6)</td>
</tr>
<tr>
<td>21–30</td>
<td>11 (2)</td>
<td>993 (4)</td>
</tr>
<tr>
<td>31–40</td>
<td>18 (3)</td>
<td>893 (4)</td>
</tr>
<tr>
<td>41–50</td>
<td>23 (3)</td>
<td>846 (4)</td>
</tr>
<tr>
<td>51–60</td>
<td>34 (5)</td>
<td>957 (4)</td>
</tr>
<tr>
<td>61–70</td>
<td>41 (6)</td>
<td>1647 (7)</td>
</tr>
<tr>
<td>71–78</td>
<td>240 (37)</td>
<td>6646 (29)</td>
</tr>
<tr>
<td>79</td>
<td>208 (32)</td>
<td>3605 (16)</td>
</tr>
</tbody>
</table>

* The 984 genes analyzed in the MPSS data set.
* All 17,849 Arabidopsis genes in the MPSS data set.
* The 650 genes analyzed in the ADA data set.
* All 22,746 Arabidopsis genes in the ADA data set.
Divergence in Expression as a Function of Sequence Divergence

In the analyses reported above, the 2 members of a gene pair were treated as 1 unit. However, we are particularly interested in divergence between members of a duplicated pair.

The normalized Manhattan distance, \( d_N \), is one measure of divergence that compares differences in the relative abundance of 2 genes across a set of libraries (even when 2 genes have very different baseline expression levels). In both data sets, we found no relationship between \( d_N \) and \( K_S \) either for the combined data or for each duplication type separately (results not shown) but did find a highly significant positive relationship between \( d_N \) and \( K_A \) independent of duplication type (fig. 4). For MPSS, the linear regression equation was \( d_N = 0.34 + 0.67 \cdot K_A \) \((P < 2.7 \times 10^{-6})\), whereas for ADA the equation was \( d_N = 0.19 + 0.90 \cdot K_A \) \((P < 7.2 \times 10^{-8})\). It is noteworthy that the intercepts of both these equations were significantly greater than 0 \((P < 2 \times 10^{-16} \text{ for MPSS}, P < 1 \times 10^{-12} \text{ for ADA})\), indicating that even pairs identical in amino acid sequence would be predicted to have quite distinct expression patterns, especially in the MPSS data set. No significant interaction with duplication type was observed in the MPSS data but there was a significant interaction in the ADA data. Dispersed pairs showed a positive relationship \( (d_N = 0.19 + 0.93 \cdot K_A, P < 2 \times 10^{-6}) \) with a significantly positive intercept \((P < 2 \times 10^{-6})\) but segmental and tandem duplicates did not.

An alternative measure is the Pearson product-moment correlation \( (r_P) \), which summarizes the similarity in relative levels of expression among libraries. As with \( d_N \), we found neither significant overall relationship between \( r_P \) and \( K_S \) nor any interaction with duplication type.
in either data set. For $K_A$, the slope of the relationship for all duplication types combined and the interaction among duplication types were not significantly different from 0 in the MPSS data set, nor was the overall slope significantly non-zero in the ADA data set. However, there was a significant interaction and a strong negative relationship between $r_P$ and $K_A$ among segmental pairs in the ADA data set (log$(1 + r_P)/(1 - r_P)) = 1.9 - 6.5 \cdot K_A, \quad P = 4.7 \times 10^{-4}$, fig. 5).

Overlap in Specificity Declines with $K_A$

An alternative way to measure expression divergence is to quantify the overlap in expression by counting the number of libraries in which both duplicates are expressed relative to all the libraries in which either are expressed. This measure, which we refer to as $I_L$ for "library identity," varies from 0 for duplicates with completely nonoverlapping expression patterns to 1 for duplicates expressed in all the same libraries.

Taking 5 TPM as the threshold for expression of a gene in a given MPSS library, the distribution of $I_L$ was strongly bimodal. Approximately 32% of the pairs shared no libraries and 27% shared all libraries. There were considerably more true pairs with $I_L = 1$ than among shuffled pairs (19.7%). We were unable to find an optimal transformation
for $I_L$ that yielded normally distributed residuals, so statistical tests must be treated with caution. Nonetheless, we found no significant overall relationship between $I_L$ and $K_S$ and found a significant overall negative relationship with $K_A$ (fig. 6A, $I_L = 0.63 - 1.12 \cdot K_A, P = 7.85 \times 10^{-08}$); there were no interactions with duplication type. In the ADA data set, the majority of pairs were expressed in nearly the same set of libraries (55% have $I_L > 0.9$ and 23% have $I_L = 1$). Only 11 pairs (3%) have $I_L = 0$. There is no significant relationship between $I_L$ and $K_S$. However, a significant negative relationship between $I_L$ and $K_A$ is observed (fig. 6B, $I_L = 0.92 - 1.22 \cdot K_A, P < 4.0 \times 10^{-09}$), particularly among dispersed duplications ($I_L = 0.91 - 1.26 \cdot K_A, P < 5.9 \times 10^{-06}$). The regression equation for the ADA data set predicts that $I_L$ would be approximately equal to 1 for $K_A = 0$, indicating that pairs identical in amino acid sequence would be expressed in a nearly identical set of libraries.

Asymmetry of Expression between Duplicates

The model of duplicate gene preservation by subfunctionalization by Force et al. (1999) predicts complementary expression patterns following duplication, especially in young duplications. Although the presence of one copy and the absence of another from a particular library (as measured above) may reflect an extreme form of specialization within a gene pair, a more subtle specialization may be detected as a complementary expression pattern in which each copy has higher expression in at least 1 library. Alternatively, the copies may not differ significantly among the libraries or may show an asymmetric pattern in which one copy predominates in all libraries that differ between the copies.

Surprisingly, in the MPSS data set, we found that most pairs had an asymmetric pattern of expression across the libraries. In over 70% of pairs, more than 2 libraries differed between the gene copies and one of the copies had higher expression in all of them (table 2A). The distribution varied somewhat between duplication types. Young ($K_S \leq 0.5$) dispersed and segmental duplicates contained the most such asymmetric pairs.

One concern with the MPSS data set is the small number of libraries and the coarse sampling of tissues for each. It is conceivable that more complementary pairs would emerge if more, or more finely sampled, libraries were examined. However, in the ADA data set, which contains 79 different libraries, we find evidence of a similar asymmetry of expression between copies (table 2B). Nearly 70% of duplicate pairs demonstrated asymmetric expression across 10 or more libraries. In contrast, a complementary expression pattern was observed in only ~10% of the pairs. Older dispersed pairs were more likely than young ones to show such complementary expression, whereas segmental pairs had the lowest level of complementarity.
The loss of functionality in 1 gene of the pair could lead to rapid asymmetric-like changes in expression level with little selective constraint. Treating size as a proxy for function, we found that 10% of our gene pairs differ in size by more than 25% (relative to the smaller of the pair). Further, these pairs show no distribution bias among duplication categories and are more likely to have complementary or equivalent expression patterns rather than asymmetric expression patterns.

Discussion

Rapid Expression Divergence among Duplicate Pairs

Our study highlights a number of striking patterns in the divergence of expression profiles between duplicated gene pairs in Arabidopsis. First, it is clear that many pairs have substantially diverged expression patterns shortly after “birth” as evidenced by the dissimilarity between pairs that are nearly identical in sequence. Previous studies in yeast, humans, and Arabidopsis have also suggested a rapid phase of initial divergence between duplicates (Gu et al. 2002; Makova and Li 2003; Blanc and Wolfe 2004; Gu et al. 2005; Yang et al. 2005; Casneuf et al. 2006). At the same time, however, the variability in the extent of expression divergence among very young pairs is quite high, as has also been seen in these other systems. One way in which rapid divergence could occur is if, as a result of incomplete duplication, the cis-regulatory sequences of the 2 genes are initially structurally dissimilar (Katju and Lynch 2003). This would only apply to tandem and dispersed duplicates; segmental pairs are structurally identical, at least initially. But the relationship between cis-regulatory change and expression divergence is highly idiosyncratic. Haberer et al. (2004) noted that tandem and segmental duplicate gene pairs had divergent expression in Arabidopsis even when they shared many similar cis-regulatory sequences and suggested that changes to a small fraction of cis-elements could be sufficient for neofunctionalization or subfunctionalization.

In addition, epigenetic differences between duplicates that are not apparent in our data set may contribute to rapid expression differentiation. For example, organ-specific differentiation of duplicate gene expression immediately following polyploidy has been observed in cotton (Adams et al. 2003), as has silencing of polyploid-derived duplicates due to hypermethylation in Arabidopsis polyploids (Wang et al. 2004). In fact, a number of epigenetic mechanisms are known that could potentially differentiate newly arisen, structurally identical, duplicates (Matzke et al. 2003; Rapp and Wendel 2005; Chen and Ni 2006). Such epigenetic effects on expression divergence could apply to all 3 classes of duplicates and lead to rapid subfunctionalization (Rodin and Riggs 2003) or nonfunctionalization.

Despite the overall pattern of rapid expression divergence, there is strong similarity in the expression profiles of some duplicate pairs that, on the basis of their sequence similarity, appear to be rather old. Some theoretical models, such as those of Lynch et al. (2001), predict that duplicate genes with redundant functions should resolve to become

<table>
<thead>
<tr>
<th>Table 2</th>
<th>The Number of Gene Pairs with Complementary Versus Asymmetric Expression Divergencea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Complementaryb</td>
</tr>
<tr>
<td>A. MPSS</td>
<td></td>
</tr>
<tr>
<td>Dispersed</td>
<td>Youngf</td>
</tr>
<tr>
<td></td>
<td>Oldg</td>
</tr>
<tr>
<td>Tandem</td>
<td>Youngf</td>
</tr>
<tr>
<td></td>
<td>Oldg</td>
</tr>
<tr>
<td>Segmental</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>B. ADA</td>
<td></td>
</tr>
<tr>
<td>Dispersed</td>
<td>Youngf</td>
</tr>
<tr>
<td></td>
<td>Oldg</td>
</tr>
<tr>
<td>Tandem</td>
<td>Youngf</td>
</tr>
<tr>
<td></td>
<td>Oldg</td>
</tr>
<tr>
<td>Segmental</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The percentage among all pairs in a given class and age is shown below each count.

Each copy had higher expression in at least 1 library.

One duplicate had higher expression in all libraries that differed and at least 2 libraries differed.

Only 1 library differed between duplicates.

No libraries differed between duplicates.

a \( K < 0.5 \)

b \( K > 0.5 \)
nonredundant within several million generations. Although expression in the same library does not constitute strong evidence for functional redundancy, a number of functional studies have raised the possibility that some duplicate pairs generated during the most recent polyploidization event still have not fully lost the overlap in their functionality (e.g., Pinyopich et al. 2003; Azevedo et al. 2006). Nowak et al. (1997) and Wagner (1999) have proposed models to explain the remarkable evolutionary persistence of redundant functionality between some duplicates; the driving force behind these models is selection for the production of offspring that do not suffer deleterious mutations in nonredundant genes.

Covariation of Expression Divergence and Redundant Genes

Covariation of Expression Divergence and Redundant Genes

Another major finding is that we did not see a positive relationship between divergence in expression profile and the number of synonymous substitutions per site ($K_S$), but we did observe a positive relationship with the number of nonsynonymous substitutions per site ($K_A$). Most strikingly, $r_p$ was negatively correlated with $K_A$ among segmental pairs in the ADA data set (supplementary table 2, Supplementary Material online). Because the segmentally duplicated pairs in this study diverged nearly simultaneously and the sequence divergence since the most recent polyploid would dwarf any pre-existing allelic variation between the future paralogs, the relationship does not appear to be due to covariation with the age of the paralogs. Rather, there appears to be a coupling between the rate of expression divergence and the rate of nonsynonymous substitution for a gene pair, presumably through covariation in the strength of functional constraint (or positive selection) among gene pairs. Zhang et al. (2002) previously reported a wide range of $K_A:K_S$ ratios among segmentally duplicated Arabidopsis genes (from less than 0.01 to over 0.6) showing that levels of constraint clearly vary among gene pairs.

Previous suggestions for a coupling between protein and regulatory divergence have come primarily from studies of orthologous genes. Duret and Mouchiroud (2000) identified a negative correlation between $K_A$ and broad tissue expression for both protein coding and untranslated regions of human–mouse orthologs with no correlation to $K_S$. Nuzhdin et al. (2004), studying orthologs between *Drosophila melanogaster* and *Drosophila simulans*, also found a correlation between $K_A$ and expression divergence but no correlation to $K_S$. In nematodes, Castillo-Davis et al. (2004) reported a correlation between protein and cis-regulatory sequence evolution in *C. elegans* and *C. briggsae* orthologs but, contrary to our findings, found no such relationship among paralogs.

Khaitovich et al. (2004) suggested, from an analysis of mammalian expression data, that expression patterns diverged as if they were, on the whole, neutral traits. Because neutrality does not preclude variable-purifying selection, this idea is not irreconcilable with the above findings on principle. But such a neutral model would also predict a positive relationship between expression divergence and $K_S$, as a proxy for time, among tandem and dispersed pairs, which we do not observe. Thus, our results are more consistent with covariation among duplicated pairs in the strength of positive selection acting on protein sequence and expression pattern (Shiu et al. 2006).

Our results also suggest that there may be stronger sequence conservation, both in protein sequence and in expression profile, for genes with greater expression breadth and greater average expression level. We found that both the breadth and the average level of expression declined with $K_A$, even among simultaneously duplicated pairs. This is consistent with a number of studies from mammals suggesting greater protein sequence conservation and selective constraint for genes with broader expression patterns (Zhang and Li 2004; Jordan et al. 2005).

Frequent Asymmetric Divergence between Duplicates

In both the MPSS data and the ADA data, we found pervasive expression asymmetry, where one copy had a higher expression level than the other in all the libraries in which they differed. Approximately 70% of the gene pairs in both data sets showed asymmetric divergence compared with less than 10% of gene pairs that showed complementary divergence. In a subset of these cases, the asymmetry is extreme; in approximately 15% of the MPSS gene pairs and 6% of the ADA gene pairs, one copy was not expressed in any of the libraries.

A slightly lower proportion of asymmetry is seen in tandem pairs within the MPSS data, which has a larger tandem pair sample. One possible explanation for the lower proportion may be gene conversion, which is thought to occur preferentially, though not exclusively, within tandem arrays when it occurs between genomic sequences in plants (Benovoy et al. 2005; Mondragon-Palomino and Gaut 2005).

A number of gene family studies in Arabidopsis have also reported asymmetric expression divergence. For example, Casneuf et al. (2006) found asymmetric expression divergence between duplicate genes that were translocated or part of small-scale duplications. In another case, Duarte et al. (2006) identified paralogous pairs of Type II MADS-box genes in Arabidopsis, where the expression of one paralog was significantly lower than the expression of its sister paralog in virtually all tissues, yet still exhibited selective constraint. Earlier, Zhao et al. (2003) reported consistently higher expression for some genes within the SKP-1–like gene family in Arabidopsis across multiple tissues when compared with their sister genes.

It is not possible to exclude the option that some of the apparently hypofunctional copies may, in fact, be functional at very low concentrations or expressed at higher levels under conditions or in tissues that have yet to be sampled (e.g., see Adams et al. 2003). For instance, if the two copies were expressed in different cells within an organ, or even in different compartments within a cell, it would not be apparent in a microarray experiment using whole-organ RNA extracts. The duplicated gene pair GMD1 and GMD2 (AT5G66280 and AT3G51160) provides an example of such a case. These genes produce variants of GDP-4-mannose 4,6-dehydratase in the cell wall. GMD2 is expressed nearly ubiquitously in Arabidopsis, excluding the root tip, whereas GMD1 is expressed almost exclusively in the root.
Conant and Wagner (2002) found that the cis-regulatory structure leads to functional novelty. Increased over time, even though the number of regulatory sequences shared between duplicate genes decreased. Papp et al. (2003b) found that the number of cis-regulatory sequences in knockout mutants of the 2 copies. GMD1 is more highly expressed in this specific tissue and, root tip (Bonin et al. 2003). Presumably, RNA sampled from the root tip, rather than the whole root, would demonstrate that GMD1 is more highly expressed in this specific tissue and, thus, the pair would be seen to have a complementary, rather than asymmetric, pattern of expression divergence.

Despite this caveat, asymmetric functional divergence does appear to be a common feature of duplicated genes (e.g., Ferris and Whitt 1979). Wagner (2002) reported that duplicated pairs in yeast show asymmetry not only in the number of stress conditions in which each copy is expressed but also in the number of interacting partners in experimental protein-interaction data and in the number of differentially expressed genes in knockout mutants of the 2 copies. Papp et al. (2003b) found that the number of cis-regulatory sequences shared between duplicate genes decreased over time, even though the number of regulatory sequences remained relatively constant and that the change in the cis-regulatory structure leads to functional novelty. Conant and Wagner (2002) found that the $K_A/K_S$ ratio was related to the degree of asymmetry between duplicated pairs in yeast, indicating that asymmetry may be driven by relaxed functional constraint or positive selection acting on one copy. Alternatively, purifying selection may be at work on the slower evolving copy as compared with the ancestral gene (Kim and Yi 2006). Recent studies of yeast and humans have identified a pattern of network asymmetry among duplicate genes, where 1 gene in the pair acquires more coexpressed gene partners than the other (Chung et al. 2006; Conant and Wolfe 2006).

The frequency of asymmetric, and the rarity of complementary, expression divergence suggests that the subfunctionalization model (Force et al. 1999) is not adequate to describe overall patterns of divergence. Nonetheless, the footprint of subfunctionalization may still be present if it is responsible for initial preservation of duplicates (Rastogi and Liberles 2005). The results (table 2) do not bear out the prediction of a higher proportion of young pairs with complementary expression patterns, though we do observe a nonsignificant trend of greater complementarity among tandem than dispersed duplicates, as is weakly predicted by the model (Lynch et al. 2001). Furthermore, despite its overall rarity, there are some striking examples of complementarity among the gene pairs examined. One such example is the duplicate pair AtMST1 and AtMST2 (AT1G79230 and AT1G16460), 2 mercaptopyruvate sulfurtransferases that not only show a complementary expression pattern but also have been found to differ in their subcellular localization to the mitochondria and cytoplasm, respectively (Nakamura et al. 2000).

**Differences among Modes of Duplication**

To what extent does the mode of duplication affect patterns of expression divergence? Because only a small portion of dispersed duplicates have $K_S$ values that match the relatively young tandem duplicates, comparing the non-polyploid, and given the problems with $K_S$ as a proxy for time (e.g., Zhang et al. 2002), comparing the non-polyploid tandem and dispersed duplicates is difficult. However, by comparing segmental duplicates with nonsegmental duplicates, we found that gene pairs in which both members were expressed were somewhat more frequent among segmental duplicates. A number of differences have also been observed in the literature. In general, segmental pairs tend to have longer retention times (Otto and Yong 2002). In yeast, the gene pairs retained from a genome duplication event are enriched in members of multisubunit protein complexes (Papp et al. 2003a) and have higher than average levels of transcriptional activity (Seoighe and Wolfe 1999). A number of recent studies have reported that the genes retained from the most recent polyploidy event in Arabidopsis (40–80 MYA) are enriched in signal transduction genes and transcription factors (Blanc and Wolfe 2004; Maere et al. 2005). Schranz and Mitchell-Olds (2006) report a correlation in the retention probability of genes following the most recent, and independent, polyploidy events in Arabidopsis and Cleome. Those genes that are not lost may then be under stronger purifying selection. Chapman et al. (2006) report that polymorphisms in the retained segmentally duplicated genes of Arabidopsis and rice tend to have fewer radical amino acid substitutions. The retained genes in Arabidopsis may also have a tendency to be those that, on average, are more dosage sensitive (Freeling and Thomas 2006), consistent with the balance hypothesis (e.g., Birchler et al. 2003), which predicts that genes retained from large-scale duplication events are more likely to be those belonging to stoichiometrically sensitive pathways.

**Assessment of MPSS and Affymetrix Data**

We found that the 2 different expression platforms, MPSS and Affymetrix, had complementary strengths and weaknesses for the study of expression divergence in duplicated genes. The chief advantage of MPSS and other signature sequencing–based platforms is the ability to distinguish the signatures of duplicate copies that would cross-hybridize on a microarray. The potential for cross-hybridization resulted in the loss of one-third of the duplicated gene pairs from the ADA data set. Nonetheless, the breadth of tissues and conditions assayed in the ADA data set provides us greater assurance that the asymmetric patterns of divergence within gene pairs were not an artifact of limited tissue sampling in the MPSS data. In MPSS, the coarseness of tissue sampling is necessitated by the expense of the technology, though this problem may be mitigated by more recent technologies developed for short-read sequencing. Recently developed methods, recently developed methods offer even more fine-grained analysis of expression levels among cell types (Birnbaum et al. 2003). The results presented here motivate further phylogenetic study of gene families using experimental data that can provide greater resolution for characterizing patterns of functional divergence between duplicated genes (e.g., Nakhamchik et al. 2004).

**Supplementary Material**

Supplementary tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe. oxfordjournals.org/).
Supplementary table 1 lists all the gene pairs used in this study along with all expression measurements (where applicable). Supplementary table 2 lists all P values for all statistically significant tests.

Acknowledgments

E.W.G. is supported by a Seeding Postdoctoral Innovators in Science and Education fellowship from the National Institute of General Medical Sciences (GM 00678). This work was also supported by National Science Foundation grants DBI-0027314 to T.J.V. and DBI-0110528 to B.C.M. T.J.V. acknowledges R. Moore, M. Rausher, and J. Willis for useful suggestions.

Literature Cited


